Synthese (2018) 195:1077–1101
https://doi.org/10.1007/s11229-016-1227-5

S.I.: MINIMALISM ABOUT TRUTH

Minimalism and the generalisation problem: on Horwich's second solution

Cezary Cieśliński[1]

Received: 9 January 2016 / Accepted: 18 September 2016 / Published online: 28 September 2016

Abstract Disquotational theories of truth are often criticised for being too weak to prove interesting generalisations about truth. In this paper we will propose a certain formal theory to serve as a framework for a solution of the generalisation problem. In contrast with Horwich's original proposal, our framework will eschew psychological notions altogether, replacing them with the epistemic notion of believability. The aim will be to explain why someone who accepts a given disquotational truth theory Th, should also accept various generalisations not provable in Th. The strategy will consist of the development of an axiomatic theory of believability, one permitting us to show how to derive the believability of generalisations from basic axioms that characterise the believability predicate, together with the information that Th is a theory of truth that we accept.

Keywords Truth · Minimalism · Generalisation problem

1 Horwichian MT and the generalisation problem

According to Horwich's minimalism (see Horwich 1999), all the facts about truth can be explained on the basis of the so-called 'minimal theory' (MT). The axioms of this theory are the instances of the following disquotational T-schema:

$$\text{(T)} \quad < p > \text{is true iff } p$$

where the expression '$< p >$' reads 'the proposition that p'. Horwich claims that the minimal theory fully characterises the content of the notion of truth. Moreover,

✉ Cezary Cieśliński
c.cieslinski@uw.edu.pl

[1] Institute of Philosophy, University of Warsaw, Warsaw, Poland

our understanding of this notion consists of our disposition to accept every (non-paradoxical) instance of (T). The final upshot is that the concept of truth becomes light and unproblematic, devoid of any deep nature for philosophers then to uncover.

One of the main concerns for the adherent of Horwichian minimalism is the (so-called) generalisation problem. How can the minimalist account for generalities involving the notion of truth? Consider, for example, the following statements:

(1) Every proposition of the form '$\varphi \rightarrow \varphi$' is true;
(2) For every φ, the negation of φ is true iff φ is not true;
(3) Every theorem of S is true (where S is some theory which we accept).

Objections have been made that Horwich's minimal theory is too weak to prove such generalisations (cf. Gupta 1993). The validity of this charge is not entirely clear, with the main stumbling block being that Horwich has never precisely delineated the collection of axioms of MT. Hence it is not possible to give an exact assessment of their truth-theoretic strength. Nevertheless, it is instructive in this context to see how effective disquotational theories can be in proving general statements of the envisaged sort. On the one hand, it is a well-known fact that some disquotational theories are weak in this respect. As an illustration, let L_T be the language obtained by extending L_{PA} (the language of Peano arithmetic) with a new one-place predicate '$T(x)$'. The expressions '$Sent_{L_{PA}}$' and '$Sent_{L_T}$' will be used to denote sentences of (respectively) L_{PA} and L_T. Let $Ind(L_T)$ be the set of all substitutions of the schema of induction '$\big(\varphi(0) \wedge \forall x\,[\varphi(x) \rightarrow \varphi(x+1)]\big) \rightarrow \forall x \varphi(x)$' by formulas of L_T. We define:

Definition 1

- $TB = PA \cup \{T(\ulcorner \varphi \urcorner) \equiv \varphi : \varphi \in Sent_{L_{PA}}\} \cup Ind(L_T)$,
- $UTB = PA \cup \{\forall x_1 \ldots x_n [T(\ulcorner \varphi(x_1 \ldots x_n) \urcorner) \equiv \varphi(x_1 \ldots x_n)] : \varphi \in L_{PA}\} \cup Ind(L_T)$,
- Theories like TB and UTB but with arithmetical induction only will be denoted (respectively) as TB^- and UTB^-.

UTB (hence also TB) is a disquotational truth theory quite weak in proving truth-theoretic generalisations. This is the content of the following theorem.

Theorem 2 *For every arithmetical formula $\varphi(x)$, if $UTB \vdash \forall x[\varphi(x) \rightarrow T(x)]$, then there is a natural number n such that $PA \vdash \forall x[\varphi(x) \rightarrow Tr_n(x)]$.*[1]

It immediately follows that (1) is not derivable in UTB, being that formulas of the form '$\varphi \rightarrow \varphi$' have an arbitrarily large complexity. It also follows that for *no* theory S, $UTB \vdash \forall \psi[Pr_S(\psi) \rightarrow T(\psi)]$ (with '$Pr_S(\psi)$' being an arithmetical formula with the natural reading 'ψ is provable in S'). The reason here is the same as before; namely, that the syntactical complexity of theorems of S will be arbitrarily large. A

[1] For the proof, see Halbach (2001, p. 1960). The expression '$Tr_n(x)$' is an arithmetical truth predicate for formulas of complexity not larger than n. The notion of complexity of a formula can be defined in various ways (for example, it can be characterised as the height of the syntactic tree of a formula). We omit the details, as they are not crucial here.

different argument shows that (2)—the compositional principle for negation—is not provable in UTB either.[2]

On the other hand, it is not the case that all disquotational theories are truth-theoretically weak. As soon as we drop the typing restrictions, the situation changes drastically, as witnessed by the following observation due to McGee (1992).

Theorem 3 *Let PAT be Peano arithmetic formulated in L_T. Let φ be an arbitrary sentence of L_T. Then there is a sentence ψ of L_T such that $PAT \vdash \varphi \equiv (T(\psi) \equiv \psi)$.*

In other words, every L_T sentence is provably (in PAT) equivalent to some substitution of Tarski's disquotational schema. In particular, this includes sentences of L_T corresponding to (1)–(3). Therefore (1)–(3) will be provable in some untyped disquotational truth theories.

At best, McGee's result shows that disquotational theories are not doomed at the start: it is theoretically possible for some set of well-motivated disquotational axioms to be truth-theoretically strong. However, it still remains unclear whether it is anything more than a mere theoretical possibility. The crucial question remains: Is there any one set of disquotational axioms which is both well-motivated and truth-theoretically strong? It is worth observing that the untyped disquotational theories which were actually proposed in the literature (with some philosophical motivation offered) do not fare well in this respect.[3]

In effect, the minimalist owes us an answer to the question of why—if at all—we are entitled or perhaps obliged to accept various generalisations involving the notion of truth. Does his disquotational truth theory prove generalisations such as (1)–(3)? And, if it does not, how can it help us to arrive at them? In a nutshell, this is the generalisation problem.

A useful way of framing the challenge has been suggested by Ketland (2005), who introduced the concept of conditional epistemic obligation. Ketland starts with the intuition that if we accept some base arithmetical theory S (formulated in L_{PA}), then we are obliged to accept various further statements, possibly unprovable in S itself. Here Ketland's emphasis is on reflection principles; the relevant definition is introduced below. The acronyms (GR), (UR) and (LR) stand for global, uniform and local reflection respectively.

Definition 4

(GR) $\forall \psi \in L_{PA}\, [Pr_S(\psi) \rightarrow T(\psi)]$.
(UR) $\forall x\, [Pr_S(\psi(x))) \rightarrow \psi(x)]$, for all $\psi(x) \in L_{PA}$.
(LR) $Pr_S(\psi) \rightarrow \psi$, for all $\psi \in Sent_{L_{PA}}$.

[2] The simplest proof known to me uses compactness: given a finite subset Z of axioms of UTB, a model of Z can be built which does not satisfy the compositional principle for negation. Hence, adding the negation of (2) to UTB produces a consistent theory.

[3] PTB and $PUTB$ are examples of such untyped disquotational theories. Their axioms contain substitutions of disquotational schemas (the local or the uniform) by *positive* formulas—formulas of L_T in which every occurrence of 'T' lies in the scope of even number of negations. However, it is known that none of them proves compositional principles for truth. For more information about these theories, see Halbach (2009) and Cieśliński (2011).

According to Ketland, when accepting S, we are epistemically obliged to accept reflection principles for S in all three versions. Since none of the reflection principles is provable in S,[4] the natural explanation of our conditional epistemic obligation suggests itself: namely, that all three principles become theorems as soon as appropriate truth axioms are added. However, the prospects for finding well-motivated disquotational axioms producing this effect look dim. Therefore, the disquotationalist faces the dilemma. Paraphrasing Ketland, either he strengthens his axioms, thereby rejecting disquotationalism, or he offers some non-truth-theoretic analysis of the conditional epistemic obligation.[5]

Indeed, it is my opinion that Ketland's concept of conditional epistemic obligation can be fruitfully applied not just to reflection principles, but also to the compositional principles governing the behaviour of the truth predicate. Consider TB as a starting point. For every arithmetical sentence ψ, we can easily establish in TB that $T(\neg\psi) \equiv \neg T(\psi)$. The recognition of this fact carries a conditional epistemic obligation: given that we accept TB, we should also accept the general compositional principle for negation (restricted, in this context, to arithmetical sentences only), even though—as it happens—it is not provable in TB itself. How can the disquotationalist account for this without compromising his philosophical standpoint? That is the question.[6]

In recent years Paul Horwich has made two attempts to deal with the challenge. They will be presented and briefly discussed in the next section.

2 Horwich's two solutions

First attempt In Horwich (1999) an attempt is made to strengthen the minimal theory in such a way that it proves by itself the desired generalisations. This strengthening involves modifying the proof techniques available to us in MT. In Horwich's words:

> It is plausible [. . .] that there is a truth-preserving rule of inference that can take us from set of premises attributing to each proposition some property F, to the conclusion that all propositions have F. (Horwich 1999, p. 137)

[4] Both (UR) and (LR) permit us to prove the consistency of S and, as such, are unprovable in S by Gödel's second incompleteness theorem. (GR) is formulated in L_T, not in L_{PA}, but even if we add the truth predicate and extend S with some natural disquotational truth axioms, the chances are high that (GR) will remain unprovable.

[5] Cf. Ketland (2005, p. 80). Although Ketland's discussion concerns conservative truth theories, in my opinion his remarks apply just as well to disquotational theories of truth in general, not necessarily conservative ones.

[6] Admittedly, in other places Ketland formulates his criticism in a different manner. Thus, arguing against Tennant, he writes: "Part of the point of the articles by Feferman, Shapiro and myself was to show how to prove reflection principles [...] As far as I can see, in the absence of the sort of truth-theoretic justification given by Feferman, Shapiro and myself, Tennant's proposal is that the deflationist may assume these principles without argument." (Ketland 2005, p. 85) Here the emphasis is on *justifying* the independent sentences (namely, reflection principles), not on explaining of why they should be accepted. In this paper I am not going to consider this quite different version of the anti-deflationary argument. Let me say only that I do not consider it successful, mainly due to the serious doubts concerning the justificatory value of truth-theoretic proofs. See Cieśliński (2015, p. 81ff) for more in this direction.

It seems that Horwich proposes the introduction of a new rule, very similar to the well-known ω-rule applied in arithmetical contexts. Assuming that we accept all the sentences obtained from $\varphi(x)$ by substituting an arbitrary numeral for the variable x, the ω-rule permits us to accept the general statement '$\forall x \varphi(x)$'. The quoted passage hints at a similar strategy. If for each proposition φ, $F(\varphi)$ can be derived in our theory, then we are entitled to conclude that $\forall \varphi F(\varphi)$.

I will not discuss this idea in detail, referring the reader to Raatikainen (2005) for a convincing criticism. The main thrust of Raatikeinen's remarks is that any rule invoked to solve the generalisation problem should be *practical*. In other words, there does not seem to be much point in generalisations being provable in a given theory of truth if we, as human beings, are never able to produce such proofs. After all, we somehow *do* reach generalisations about truth and any adequate explanatory account should take this fact into consideration. Unfortunately, the system with the ω-rule does not satisfy this basic feasibility condition. The rule in question requires infinitely many premises and for this reason its practical utility is close to null.[7] Indeed, I am inclined to think that it is a very serious worry.

Second attempt Horwich's second solution was originally proposed in Horwich (2001) and elaborated on in Horwich (2010). Unlike in the previous case, the current proposal involves leaving the proof machinery of MT intact (it remains thoroughly classical); the idea is just to use it together with a certain additional premise. Horwich emphasises that, apart from MT, the minimalist is permitted to uses additional 'truth-free' assumptions in his explanations. For example, we can explain why we accept '<Elephants have trunks> is true' as soon as we enlarge MT with a truth-free assumption 'Elephants have trunks'. Here is the final answer: we accept that 'Elephants have trunks' is true because we believe that elephants have trunks and we accept an appropriate disquotational axiom of MT.

In an attempt to generalise this strategy, Horwich proposes the following truth-free assumption:

(A) Whenever someone is disposed to accept, for any proposition of structural type F, that it is G (and to do so for uniform reasons) then he will be disposed to accept that every F-proposition is G. (Horwich 2010, p. 45)

With this assumption at hand, Horwich promises to explain why we are inclined to accept generalisations of the (1)–(3) type. As an example, we present the Horwichian explanation below for (1).

Explanation 5

(P_1) *For every proposition of structural type '$\varphi \rightarrow \varphi$', we are disposed to accept that it is true (and we do it for uniform reasons).*

(P_2) *If P_1, then we will be disposed to accept that every proposition of structural type '$\varphi \rightarrow \varphi$' is true.*

[7] Cf. Raatikainen (2005, p. 176): "The ω-rule has its uses in theoretical contexts, but because of its infinitary nature, it is not a rule of inference in the ordinary sense. That is, the usual rules of inference are decidable relations between (conclusion) formulas and finite sets of (premiss) formulas. This is not so with the ω-rule. It requires that one can, so to say, have in mind and check infinitely many premisses, and then draw a conclusion. Consequently, we finite human beings are never in a position to apply the ω-rule".

Conclusion We will be disposed to accept that every proposition of structural type '$\varphi \rightarrow \varphi$' is true.

Clearly, the reasoning is logically valid. Indeed, we can assume that premise (P_1) describes us as users of a given disquotational theory of truth; it is also easy to observe that P_2 is an instance of (A). In effect, we obtain the explanation of our acceptance of '$\forall\varphi\, T(\varphi \rightarrow \varphi)$'.

Still, some critics remained unconvinced. In particular, Bradley Armour-Garb was dissatisfied with premise P_2. In his own words:

> One will not be disposed to accept (the proposition) that all F-propositions are G, from the fact that, for any F-proposition, she is disposed to accept that it is G [...], unless she is aware of the fact that, for any F-proposition, she is disposed to accept that it is G. (Armour-Garb 2010, p. 699)

Armour-Garb's reservation seems fair indeed. However, as he notes himself, it is possible to take this objection into account, which generates the following modified version of Explanation 5:

Explanation 6

S_1 *For every proposition of structural type '$\varphi \rightarrow \varphi$', we are disposed to accept that it is true.*
S_2 *We are aware that S_1.*
S_3 *If S_1 and S_2, then we will be disposed to accept that every proposition of structural type '$\varphi \rightarrow \varphi$' is true.*

Conclusion We will be disposed to accept that every proposition of structural type '$\varphi \rightarrow \varphi$' is true.

Nevertheless, Armour-Garb is dissatisfied with S_2. He asks: "What is it for one to be aware of such a fact"? He then answers:

> Here is a plausible answer: for one to be aware of the fact that, for every F-proposition, she is disposed to accept that it is true is for that person to be aware of the fact that she is disposed to accept that every F-proposition is true. (Armour-Garb 2010, p. 700)

If this is so, then S_2 simply means 'we are aware that the conclusion holds' and for this reason Armour-Garb accuses Explanation 6 of being viciously circular. We just cannot explain our disposition to accept a general sentence by citing our awareness that we have such a disposition.

Still, such a dismissal of Horwichian explanations seems to me too hasty, since there are other possible interpretations of S_2 which should be taken into account. The next section contains an initial sketch of what seems to me to be a more promising strategy.

3 Horwichian explanations reconsidered

We will propose here a certain reconstruction of Horwichian explanations. To enhance clarity, they will be presented in a very restricted, arithmetical framework. For starters, we are going to assume that the disquotational theory TB^- is our preferred theory of truth for the language of arithmetic. The obvious question then arises of why we are inclined to accept such general statements as (1).[8] A Horwichian explanation of our acceptance of '$\forall\psi \in Sent_{L_{PA}}\ T(\psi \rightarrow \psi)$' will be presented below. The explanation is carried out in a metatheory S about which we will make the following stipulations.

(a) The language of S contains expressions 'we are aware that …' and 'we are disposed to accept …', predicated of sentences of L_S.
(b) S contains Peano arithmetic.
(c) S contains the information that TB^- is our theory. That is, we are disposed to accept sentences in awareness that they are theorems of TB^-.
(d) S contains an axiom stating that if we are aware that (for every x, $\varphi(x)$), then for every x, we are aware that $\varphi(x)$.[9]
(e) S contains the following necessitation rule: given φ as a theorem, we are allowed to infer 'we are aware that φ'.
(f) S contains Horwich's rule: given a proof of 'we are aware that for every x, we are disposed to accept $\varphi(x)$', we are allowed to infer: 'we are disposed to accept $\forall x\varphi(x)$'.[10]

It should be noted that, as it stands, S describes us as highly idealised users of TB^-.[11] Thus, for example, condition (b) together with the closure condition (e) guarantees that for every theorem φ of PA, S will prove 'we are aware that φ'. Surely, the awareness of every single theorem of Peano arithmetic (including those never proved by anyone) is an impossibly tall order for any real-world agent, which invites the charge that any solution to the generalisation problem based on S will be as unrealistic and impractical as the recourse to ω-rule. Nonetheless, in fact the situation is not that dire at all. In the course of explaining the dispositions of the real-world agents, we can still appeal to those concrete reasonings carried out in S, which employ only the

[8] Restricting our attention to Peano arithmetic and TB^- brings both gains and losses. On the one hand, we gain clarity, since the set of axioms of TB^- (unlike that of MT) is precisely defined. On the other hand, we admittedly lose the breadth and scope of Horwich's original proposal. Indeed, Horwich discusses arbitrary *propositions*, not arithmetical sentences, and *properties*, not formulas. Here we are going to sacrifice scope for the sake of clarity. However, it is worth emphasising that if Horwichian explanations do not work in simple arithmetical contexts, then they are even more problematic when applied to propositions and properties.

[9] One should be careful, nonetheless, about the use of implication. Let us abbreviate 'I am aware that x' by '$A(x)$'. Given '$A\big(\forall x[\varphi(x) \rightarrow \psi(x)]\big)$', by stipulation (d) I can infer 'for every x, $A\big(\varphi(x) \rightarrow \psi(x)\big)$'. 'What I cannot do is to automatically infer 'for every x, if $\varphi(x)$, then $A(\psi(x))$'. Even taking this reservation into account, one could wonder what psychological reality corresponds to (d). My suggestion is that, on the assumption of a minimal logical competence of the agent, the awareness of the general fact generates something more than just a disposition to accept all the instances, namely, the explicit knowledge of a simple algorithm producing, for an arbitrary n, a derivation of $\varphi(n)$ from the general statement.

[10] This clearly corresponds to Horwich's assumption (A), even though I formulate it as an inference rule here.

[11] I am grateful to the anonymous referee for this observation.

principles *known* to the agents at a given time (ideally, principles for which the agents themselves have provided proofs).[12]

At this point let us recall Armour-Garb's question. What is it for one to be aware that φ? In my opinion, the application of (e) to the real-world agents (see the previous paragraph) provides a reasonable sufficient condition. In order to be aware that φ it is enough to *prove* φ. After proving φ in our metatheory S, we are permitted to conclude 'we are aware that φ'.

Below, we present a Horwich-style explanation of why we are inclined to accept that every arithmetical sentence of the form '$\varphi \rightarrow \varphi$' is true. The explanation proceeds as follows.

Explanation 7

(1) *For every* $\varphi \in Sent_{L_T}$*, if we are aware that* $TB^- \vdash \varphi$*, then we are disposed to accept* φ. (By (c))
(2) *For every* x, $TB^- \vdash Sent_{L_{PA}}(x) \rightarrow T(x \rightarrow x)$. (By (b), since (2) is provable already in PA)[13]
(3) *We are aware that* (2). (Necessitation, applied to (2))
(4) *For every* x*, we are aware that* $TB^- \vdash Sent_{L_{PA}}(x) \rightarrow T(x \rightarrow x)$. (From (3) by (d))
(5) *For every* x, $Sent_{L_T}(\ulcorner Sent_{L_{PA}}(x) \rightarrow T(x \rightarrow x)\urcorner)$[14] (Provable in PA)
(6) *For every* x*, we are disposed to accept:* $Sent_{L_{PA}}(x) \rightarrow T(x \rightarrow x)$. (From (1), (4) and (5), logic)
(7) *We are aware that* (6). (Necessitation, applied to (6))
(8) *We are disposed to accept: for every* x, $[Sent_{L_{PA}}(x) \rightarrow T(x \rightarrow x)]$. (From (7) by Horwich's rule)

The acceptance of various other truth-involving generalisations can be explained in an analogous manner.

How should we assess Explanation 7? All in all, I am inclined to think of it as of a step in the right direction. Nevertheless, I will formulate two critical remarks below, both of which indicate the need for a still further reformulation and amendment.

Before engaging into the criticism, let us note one peculiar trait of Explanation 7 (or of Horwichian explanations in general). The explanation of why we are disposed to accept a given statement could proceed by deriving this statement in a theory which we accept. Indeed, this would be the case of TB^- with ω rule, where various truth-theoretic generalisations become provable; this is also the case when we try to explain our acceptance of the truth of 'Elephants have trunks'.[15] However, this is not what happens here. The general statement in question, i.e. '$\forall \varphi \in L_{PA} T(\varphi \rightarrow \varphi)$', is not

[12] Alternatively, the rules of S could be modified by relativising them to agents and times.

[13] As usual in such contexts, the intended meaning is that for every x, TB^- proves the result of substituting a numeral denoting x for a free variable in the relevant formula.

[14] This notation is used here as a shorthand of: 'for every x, y, if $y = \ulcorner Sent_{L_{PA}}(x) \rightarrow T(x \rightarrow x)\urcorner$, then $Sent_{L_T}(y)$', with '$y = \ulcorner Sent_{L_{PA}}(x) \rightarrow T(x \rightarrow x)\urcorner$' abbreviating '$y$ is the result of substituting a numeral denoting x for a variable v in the expression "$Sent_{L_{PA}}(v) \rightarrow T(v \rightarrow v)$"'.

[15] 'T(Elephants have trunks)' is simply provable in a disquotational truth theory enriched with an additional assumption 'Elephants have trunks'.

derived here at all, neither in TB^- (which would be impossible, anyway), nor even in TB^- supplemented with some additional premises. What is instead derived is a statement about our *disposition* to accept the general sentence under discussion.

I will now formulate two objections against Explanation 7 and against Horwichian explanations in general.

Problem 1 Horwichian explanations are psychological. A psychological fact (namely, our disposition to accept a given sentence) is explained here in terms of our other dispositions and mental abilities. This in itself is not problematic. There is nothing wrong with psychological explanations as such. However, the trouble is that in Explanation 7 the normative element is completely lost and the following additional question arises: Is someone who accepts TB^- (or Horwich's MT) *committed* to accept additional generalisations, unprovable in TB^-? Assume for the sake of argument that we do satisfy the description from Explanation 7, entailing that, while accepting TB^-, we are also inclined to accept the general statement '$\forall\varphi \in L_{PA}T(\varphi \rightarrow \varphi)$'. But is there any reason why we *should* accept such independent sentences?

If the sentence in question was provable in a theory accepted by us, the difficulty would not be so acute. However, as we noticed, Explanation 7 does not contain a derivation of '$\forall\varphi \in L_{PA}T(\varphi \rightarrow \varphi)$' in any theory accepted by us. Why then should we accept it?[16]

Problem 2 Premise (1) requires some reformulation. It seems that an assumption to the effect that TB^- is a theory accepted by us, should be employed in Horwichian explanations. But what does it mean to accept a theory? The problem is that Premise (1) does not adequately express this content. For illustration, assume that my knowledge of the theory TB^- is very limited—that I do not know much more about TB^- apart from the fact that it is some theory. In such a case Premise (1) would be vacuously true; nevertheless, we would not say that in this situation I accept TB^-.

In view of this, later on we are going to propose an alternative approach, which preserves some essential traits of Horwichian explanations, but gets rid of psychological concepts altogether. However, we will start with remarks about the notion of accepting a theory.

4 Accepting a theory

Given that we accept a theory Th, why should we then accept various generalisations that are not provable in Th? This question is the starting point of our investigations; this is also what we consider to be the basic challenge behind the generalisation problem. In such a formulation, the notion of accepting a theory comes out as crucial. But what does it mean to accept a theory?

[16] Admittedly, the conclusion of Explanation 7 is that we are disposed to accept that $\forall\varphi \in Sent_{L_{PA}}T(\varphi \rightarrow \varphi)$. Hence, we could point out that the sentence in question does, after all, follow quite trivially from some theory accepted by us (namely, from the theory containing this very sentence). However, this would be a moot point. The question still remains how we arrived at such a theory and, more importantly, why we *should* embrace it.

For starters, we are going to consider (and reject) a few candidates for the role of an explication of 'I accept Th'. Our criterion of assessment of the explications will be twofold. First, we are going to reject those explications which yield clearly implausible consequences.[17] Second, we will also reject those explications which are too strong (in a sense to be explained below). Here is our initial list of options. 'I accept Th' could mean that:

(a) For any sentence φ, if I believe that φ has a proof in Th, then I am ready to accept φ.
(b) For any sentence φ, if I believed that φ has a proof in Th, then I would be ready to accept φ.
(c) I accept that all the theorems of Th are true.
(d) I accept some truth-free version of the reflection principle for Th (the local or the uniform one).

(a) is clearly inadequate for reasons which have already been indicated (see *Problem 2* above). Indeed, if I know nothing about Th, then (a) is vacuously true. Hence it would follow that I accept Th, which is hardly plausible.

(b)—a counterfactual strengthening of (a)—turns out to be inadequate as well. I submit that very few people would accept any theory Th in such a sense! For example, if I believed that Peano arithmetic proves '$0 = 1$', then I would not be ready to accept that $0 = 1$. I would reject PA instead.

Even though (c) sounds plausible in itself, it is too strong to be of much use in our present discussion. To give a simple example, one of the 'conditional epistemic obligations' (to use Ketland's phrase again) is the consistency of Th. Why is it that when accepting Th, we should accept that Th is consistent? In a way, the explication (c) makes the issue trivial. Thus, given an arithmetical theory Th it is easy to observe that any theory of truth which includes T-sentences for the arithmetical language, proves Con_{Th} (the consistency of Th) when supplemented with 'All theorems of Th are true'. So far so good—but where does it leave the disquotationalist? How can he accept Th if—as it may well happen—his disquotational truth theory for the language of Th does not permit him to prove 'all theorems of Th are true'? Indeed, if (c) was the only possible way to make sense of the notion of 'accepting a theory', I would be inclined to see it as a strong argument against disquotationalism. But, this is not the only way.

In addition, treating (c) as an explication of 'I accept Th' will be very problematic in some particularly pertinent cases, namely, when Th itself is a theory of truth. For illustration, let Th be $KF + \text{CON}$; in other words, let it be the Kripke–Feferman system supplemented with the consistency axiom '$\forall\varphi\neg(T(\varphi) \wedge T(\neg\varphi))$'.[18] What does

[17] For example, if under a given explication I 'accept' a theory which I do not accept in a normal sense of the word, this will count heavily against the proposed explication.

[18] See Reinhardt (1986) and Feferman (1991), where the theory in question was introduced; cf. also Feferman (1984). KF was meant to capture a Kripkean notion of truth and originally it was formulated in the language with two primitive predicates 'T' and 'F' for truth and falsity respectively. For the list of axioms of KF in the language with 'T' but without the falsity predicate, see Halbach (2011, pp. 200–201).

it mean to accept such a theory? Since the system $KF + \text{CON}$ proves the untruth of some of its own theorems, accepting $KF + \text{CON}$ in the sense (c)—that is, introducing the information that all theorems of $KF + \text{CON}$ are true—produces an inconsistent theory.[19] This is another reason why (c) should be deemed unsatisfactory.

We reject also (d) for similar reasons as (c)—that is, we consider it too strong for our purposes. If 'accepting Th' means accepting all the substitutions of (say) the local reflection principle for Th, then how can the disquotationalist accept Th in this sense if (as it may well happen) his disquotational truth theory does not prove all such substitutions? Indeed, one could try to argue that on any admissible sense of 'accepting', our acceptance of Th commits us to some forms of reflection for Th.[20] Still, I am inclined to view such a commitment as something to be *explained*. It is totally unhelpful to postulate it in advance, as a part of the meaning of 'I accept Th'.

At this point we have considered and rejected four explications of 'I accept Th'. So what is left?

In this paper we will work with the following explication, which is a modified version of (b).

(e) For any sentence φ, if I believed that φ has a proof in Th and I had no independent reason to disbelieve φ, then I would be ready to accept φ.

Observe that the objection raised earlier against (b) is no longer valid. Here, a mere possibility of the theory being inconsistent is no longer a problem: if I believed that '$0 = 1$' has a proof in Th, then I would *not* be ready to accept that $0 = 1$ because of independent reasons! In fact, the formulation (e) gives justice to the fact that we rarely—if ever—accept our theories unconditionally. The intuition is rather that we will stick to them as long as we believe that they do not yield false consequences. Still, if I accept Th, then given a *new* sentence φ (new in the sense that I neither accepted nor rejected φ previously), I would accept φ if I believed that it is a theorem of Th. Indeed, this is how the present story goes.

Throughout the rest of this paper I am going to treat (e) as my basic description of the content of 'I accept Th'. Let us observe that, with such a choice, the generalisation challenge remains nontrivial. For example, we still have to explain why someone accepting PA in this sense should be committed to accept statements not provable in PA (the consistency statement in particular). Alternatively, one could claim that any such additional commitments are illusory.

[19] For a sentence L such that $KF \vdash L \equiv \neg T(L)$, it is possible to show that $KF + \text{CON} \vdash L$ and therefore $KF + \text{CON} \vdash \neg T(L)$. It is then easy to observe that $KF + \text{CON}+$ 'All theorems of $KF + \text{CON}$ are true' proves $T(L)$ and hence it is inconsistent. See Halbach (2011, p. 215).

[20] Cf. Ketland's 'If one accepts a mathematical base theory S, then one is committed to accepting a number of further statements in the language of the base theory (and one of these is the Gödel sentence G)'. Ketland (2005, p. 79) In this context, the role of reflection principles is quite central: they give us 'the possibility of systematically generating larger and larger systems whose acceptability is implicit in acceptance of the starting theory. The engines for that purpose are what have come to be called reflection principles' (Feferman 1991, p. 1). These words of Feferman are quoted with approval by Ketland in the same paper.

5 An epistemic approach

5.1 Introducing believability

In the remaining part of this paper, a Horwich-style solution to the generalisation problem will be proposed, one which eschews psychological concepts altogether. Instead of 'being aware of' and 'being disposed to accept' (see Explanation 7), in our amended explanations we are going to employ a single epistemic predicate $B(x)$ ('x is believable'), predicated of sentences. The intuitive intended interpretation of $B(x)$ is 'there is a reason to accept x which is good enough to warrant rational acceptance of x, given the absence of reasons to reject x' (later on we will abbreviate this as 'there is a reason to accept x which is normally good enough' or just as 'there is a good reason to accept x'). We emphasise that this interpretation is not to be confused with 'there is a compelling reason to accept x' or with an unconditional 'x should be rationally accepted'. It corresponds rather to the weak notion of theory acceptance characterised by condition (e) on p. 11 of this paper. The initial idea is that proofs carried out in a theory that we accept are treated by us exactly as such reasons: when presented with such a proof, we accept its conclusion, unless given strong reasons to the contrary.[21] In a moment we are going to characterise the new predicate by means of axioms and rules. However, I will start with a general outline of the proposed strategy.

We begin with a very simple notion of truth, characterised by purely disquotational axioms of our truth theory Th (resembling TB^- or, more ambitiously, resembling perhaps Horwich's MT). We are convinced that our axioms fully specify the meaning of the truth predicate. Moreover, we treat the axioms as obvious, simple and epistemologically basic. Sometimes we express these convictions by slogans like 'truth is innocent' or 'truth is a light notion'.

We accept our disquotational truth theory. The notion of acceptance is employed here in the sense (e) of the previous section. In practice, given a reliable information about the existence of a proof of φ in Th, we accept φ. However, there is one additional element of the picture: it should be emphasised that in such cases we accept φ *because of* the proof in Th. Indeed, we consider theorems of Th believable. The key point is that our mathematical practice—that of accepting statements because of their proofs in Th (even in cases when we did not check the proofs by ourselves)—would be irrational without the underlying belief that proofs in Th function as reasons which are normally good enough to accept their conclusions.

In the next stage, we characterise our notion of believability by means of some basic axioms and rules. As we are going to see, this move brings important consequences. The final result is that we declare as believable various *additional* statements in the language of Th, unprovable in Th itself. In effect, our initial acceptance of disquotational theory Th, together with some basic convictions about believability, leads us to

[21] Similarly, information from a very trustworthy witness will warrant rational acceptance given the absence of reasons to reject the testimony (in particular, given the absence of contrary testimonies from other trustworthy witnesses). On the other hand, information conveyed by an unreliable witness is not believable in our intended sense: the reason in question (namely, the witness's words) does not warrant rational acceptance even in the absence of any evidence to the contrary—this is the intuition.

the recognition that, indeed, there is a reason to accept e.g. compositional truth axioms. In this way new commitments are generated. The point is that given such a reason, we behave rationally when—not aware of any good reasons to reject compositional principles—we finally accept them. This is the general outline of the story proposed here.

In order to make some of these ideas precise, we introduce the following definition:

Definition 8 Let K be an axiomatisable extension of PA in the language L_K (possibly richer than L_{PA}). Denote as $L_{K,B}$ the extension of L_K with a new one-place predicate 'B'. Let KB be a theory K formulated in the language $L_{K,B}$.[22]

- We denote as $Bel(K)^-$ the theory in the language $L_{K,B}$ which extends KB with the following axioms:

  $(A_1)\ \forall\psi \in L_{K,B}[KB \vdash \psi \rightarrow B(\psi)]$
  $(A_2)\ \forall\varphi, \psi \in L_{K,B}[\big(B(\varphi) \wedge B(\varphi \rightarrow \psi)\big) \rightarrow B(\psi)]$

  In addition, the theory $Bel(K)^-$ has the following rules of inference:

  $$\text{NEC}\ \frac{\vdash \phi}{\vdash B(\phi)} \qquad \frac{\vdash \forall x B\phi(x)}{\vdash B\big(\forall x \phi(x)\big)}\ \text{GEN}$$

- We denote as $Bel^{Con}(K)^-$ the theory which is exactly like $Bel(K)^-$, except that it contains the following additional consistency axiom:

  $(A_3)\ \forall\psi \in L_{K,B} \neg B(\psi \wedge \neg\psi)$.

- $Bel(K)$ and $Bel^{Con}(K)$ are theories which are exactly like $Bel(K)^-$ and $Bel^{Con}(K)^-$, except that they contain all the axioms of induction for formulas of $L_{K,B}$.

We will view $Bel(K)$ as a believability theory built over our initial theory K. The intended reading of '$B(x)$' ('x is believable') will manifest itself in our treatment of a proof of $B(\psi)$ in $Bel(K)$ as showing that ψ should be rationally accepted, given our acceptance of K and given that we are not aware of any independent reasons to reject ψ.[23]

Axiom (A_1) is not much more than a formalisation of the assumption hidden behind our acceptance of K. We realise that proofs in K are (for us) good reasons to accept their conclusions—that is the content. The only addition is that (A_1) also expresses our acceptance of logic in the extended language, with the new predicate 'B' (we emphasise that in the formulation of (A_1) the provability predicate for KB is used).

Axiom (A_2) expresses the following thought: if there is a good reason to accept the implication and there is a good reason to accept its antecedent, then there is a good reason to accept the consequent.

As for the rules of inference of $Bel(K)$, the intuitive validity of NEC seems to me uncontroversial. If a proof of φ in $Bel(K)$ is provided, it is simply this proof

[22] The only difference between K and KB is that KB contains logical axioms also in the language with the new predicate.

[23] As noted earlier, the last reservation is clearly forced by our weak notion (e) of accepting a theory.

itself which constitutes a good reason to accept φ, therefore in such a situation φ is believable. [24]

Rule GEN is of crucial importance. As we are going to see, it is exactly GEN which permits us to derive (in the scope of 'B') strong consequences, unprovable in K itself. In order to explain the idea behind GEN, let us compare it to the following (somewhat similar) reflection rule for the theory K[25]:

$$\frac{\vdash \forall x\, Pr_K \phi(x)}{\vdash \forall x \phi(x)} \text{ REF}$$

Let us assume that K is the theory we accept. Let us also assume that we are able to prove (perhaps in K itself) that for every x, $\phi(x)$ is a theorem of K. Why should we then accept the general statement '$\forall x\ \phi(x)$'? The answer becomes unproblematic as soon as we assume that all theorems of K are true. Indeed, after this (with just a few basic properties of the truth predicate) the conclusion of REF can be easily justified. Still, with just the disquotational notion of truth we might not have a right to such an assumption. How can we close this gap?[26]

The idea of GEN is to close the gap by introducing an additional epistemological element. Here, it should be noted that what we really do in such contexts involves some reasoning about *reasons*, not about truth. For illustration, assume that you consider the Pope an authority in matters of religious faith. You trust the Pope and the fact that a certain opinion was expressed by the Pope is treated by you as a good reason to accept this opinion. Assume in addition that you heard from a trustworthy source—possibly from the Pope himself—that every book written by a given theologian was proclaimed heretical by the Pope. You will then have a good reason to think that (for every book written by the theologian, there is a good reason to think that it contains a heresy). The idea behind GEN is that this in itself constitutes a good reason to think that every book written by the theologian contains a heresy. According to the present proposal, this is how we reason about reasons.

It is worth observing that the following transformation of GEN into the axiom misses the mark:

(Ax) $\forall \psi[\forall n B(\psi(n)) \rightarrow B(\forall n \psi(n))]$.

As a formalisation of our intuitions concerning believability, (Ax) is clearly inadequate. For every numerical instance of ψ, there may exist a good reason to accept it without there being a *uniform* reason, one covering all the cases simultaneously. In other words,

[24] Note that, unlike in the case of (A_1), NEC is the closure condition for $Bel(K)$, so it can be applied also to statements which are not theorems of KB. On the other hand, NEC by itself has not the full force of (A_1). Indeed, for every theorem φ of KB, it is enough to have NEC (without (A_1)) in order to prove $B(\varphi)$. Still, NEC by itself is not enough to derive the believability of all theorems of KB as a general statement.

[25] See Beklemishev (2005) for a more detailed discussion of the reflection rule. In particular, Beklemishev shows that REF is equivalent to (UR) over Peano arithmetic.

[26] Indeed, one might be tempted to interpret Horwich's assumption (A) as containing the proposal of enriching MT with something like REF (or maybe even with the uniform reflection principle stated in the form of an implication). However, the key question would then be what gives us the right to introduce such a rule—why should we accept its conclusion, given the acceptance of the premise?

there might still not be any good reason to believe the general statement in question. We emphasise that in GEN this particular flaw is eliminated; we require that the statement '$\forall n B(\psi(n))$' is itself provable, and that we therefore have a good, uniform reason (namely, the proof of the general sentence in $Bel(K)$) to believe of every instance of ψ that it is believable. The intuition is that such a good reason to think that all instances are believable is also a good reason to believe a general statement. It is exactly this intuition that is expressed by GEN.

Axiom (A_3) is an ugly duckling and I do not present it as a part of the minimal believability theory $Bel(K)$. In this context we should emphasise one more time that, according to the intended interpretation, believability is neither truth nor unconditional rational acceptance. Not only can false sentences be believable but there is also no automatic transition from '$B(\varphi)$' to 'φ should be rationally accepted'. Intuitively, '$B(\varphi)$' instead means that 'there is a reason to accept φ which is normally good enough' or, in other words, 'you should rationally accept φ unless you have a strong independent reason to reject φ'. Observe that this reading corresponds closely to the characterisation of the weak notion of theory acceptance adopted in this paper. Thus, the information that φ is provable in K (a theory which I accept) is normally good enough for me to accept φ. However, my acceptance of K is not unconditional and if it turned out that $K \vdash 0 = 1$, then I would *not* be ready to accept that $0 = 1$. In our framework, this would be the situation of $Bel(K)$ proving $B(0 = 1)$. I then still have a reason to accept '$0 = 1$' that is normally good enough (namely, the proof in K—a theory which I accept). However, the situation envisaged is not normal, since I am also able to prove that $B(\neg 0 = 1)$. In general, the intuition is that whenever we obtain both $B(\varphi)$ and $B(\neg\varphi)$, there is no automatic transition to the rational acceptance of statements in the scope of the believability predicate.

All in all, one might have a good reason—that is, a reason normally good enough—to accept a sentence φ (for instance, a derivation of φ from some plausible premises) while at the same time having a good reason to accept $\neg\varphi$ (say, a derivation of $\neg\varphi$ from another set of plausible premises). This is why I reject (A_3) as an axiom formalising the properties of the believability predicate. The axiom is introduced just in order to indicate that nice models (in a sense to be explained) can be provided even for full $Bel^{Con}(K)$ on the assumption that the initial theory K is 'safe' enough. Nevertheless, none of the applications of the believability theory discussed in this paper will employ (A_3).

5.2 Formal properties of $Bel(K)$

In this section we state the main formal results about the properties of $Bel(K)$. We omit the proofs, which can be found in the Appendix.

When analysing formal properties of $Bel(K)$ with the intended interpretation of 'B' in mind, the following question looms as primary. Imagine that K is an initial theory which we accept. If K is trustworthy, just how trustworthy are the statements which are, provably in $Bel(K)$, within the scope of B? In other words, if $Bel(K)$ proves the believability of a given statement, will something go wrong if we accept this statement? In order to facilitate further discussion, we introduce the following notational convention:

Definition 9 $Int_{Bel(K)} = \{\psi \in L_{K,B} : Bel(K) \vdash B(\psi)\}$

The set $Int_{Bel^{Con}(K)}$ can be defined in an analogous manner.

What we would like to know is whether the elements of $Int_{Bel(K)}$ form a nice theory. One particular interpretation of the phrase, 'a nice theory', will be considered here; that is, nice theories are interpretable in the standard model of arithmetic.[27] One of our main formal observation is that even $Int_{Bel^{Con}(K)}$ is indeed nice in this sense. This is the content of the theorem formulated below.

Theorem 10 *Let K be a theory in the language without the predicate 'B'. If N (the standard model of arithmetic) is expandable to a model N^* of K, then N^* is expandable to a model of $Int_{Bel^{Con}(K)}$.*

Having checked that $Int_{Bel(K)}$ is a nice theory, we observe that even without the axiom (A_3), the theory $Bel(TB^-)$, obtained by taking TB^- as the basic (initially accepted) theory, is already strong enough to prove the believability of various interesting generalisations unprovable in TB^-. Let '$B(CT)$' be a shorthand of 'all the compositional truth principles are believable'. It is then possible to show that:

Theorem 11 $Bel(TB^-) \vdash B(CT)$.

As a final observation, we remark that similar results can be obtained for some untyped disquotational truth theories taken as a starting point. Thus, in (Horsten and Leigh 2015) the language $L_{T,F}$ has been defined in the following manner.

Definition 12

- Terms and function symbols of $L_{T,F}$ are exactly those of L_{PA}.
- The connectives of $L_{T,F}$ are $\wedge$ and $\vee$.
- The predicates of $L_{T,F}$ are: $=$, $\neq$, T, F.
- In $L_{T,F}$ we have the quantifiers $\forall$ and $\exists$.
- The formulas of $L_{T,F}$ are built in the usual style.

What is important is that in $L_{T,F}$ we do not have a symbol for negation. We will instead be using the symmetric relation of being a dual formula. Strictly speaking, this notion is first defined for primitive symbols and then generalised to cover all formulas.

Definition 13

- '$=$' and '$\neq$' are dual predicate symbols.
- 'T' and 'F' are dual predicate symbols.
- '$\wedge$' and '$\vee$' are dual connectives.
- '$\forall$' and '$\exists$' are dual quantifiers.
- ψ and φ are dual formulas iff ψ is obtained from φ by replacing every symbol in φ with its dual.

We will use the notation φ^d whenever we want to indicate that a given formula is a dual of φ.

[27] It follows in particular that nice theories are ω-consistent.

In this language, the following basic disquotational theory can be characterised:

Definition 14

- $TFB = PA \cup \{T(\ulcorner\varphi\urcorner) \equiv \varphi : \varphi \in L_{T,F}\} \cup \{F(\ulcorner\varphi^d\urcorner) \equiv \varphi : \varphi \in L_{T,F}\} \cup Ind(L_{T,F})$.
- A theory like TFB, but with arithmetical induction only, will be denoted as TFB^-.

Let KF be a variant of Kripke–Feferman truth theory in the language $L_{T,F}$.[28] Let '$B(KF)$' be the sentence stating that all compositional axioms of KF are believable. Then it is possible to show that:

Theorem 15 $Bel(TFB^-) \vdash B(KF)$.

In this way a Horwich-style solution to the generalisation problem is vindicated. Indeed, it turns out that if we find our initial disquotational theory (for example, TB^- or TFB^-) believable, then we will also find believable various truth-theoretic generalisations, not provable in the initial theory of our choice.

5.3 How does it help?

How attractive is the proposed epistemic strategy? Does it really permit the disquotationalist to escape the generalisation problem? We believe that it does. Nevertheless, in this final section we are going to discuss some limitations of the proposed believability framework, indicating the areas with an additional work still awaiting to be done.

For starters, let us recapitulate the proposed solution.

The disquotationalist begins with accepting some disquotational theory K, which may well be truth-theoretically weak. In particular, it might happen that various compositional principles involving truth are not provable in K. Should he accept these principles? If so, why? This is the question.

Here is the proposed answer. The disquotationalist treats proofs in K as good reasons to accept their conclusions (more exactly, he treats them as reasons which are normally good enough—see the condition (e) earlier in this paper). In this realisation, he applies the believability theory to K treating it as the base. However, it transpires that $Bel(K)$ proves the believability of compositional principles. The initial acceptance of K, together with some basic convictions about believability, leads the disquotationalist to the recognition that there is indeed a good reason to accept compositional principles. Given such a good reason and without being aware of any good reason to reject these principles, he *should* accept them.

Viewed in more general terms, our endeavour has been to understand the commitments of someone who accepts a given theory K.[29] We have noticed that the answer depends on the notion of accepting a theory. After observing that some strong notions trivialise the issue, we have decided to focus on the weak notion of theory acceptance. This has forced us to adopt a notion of believability partially characterised by the axioms (A_1) and (A_2) but without the consistency axiom (A_3). We have argued that

[28] See (Horsten and Leigh 2015) for the full list of axioms.

[29] For a classical analysis of such commitments, we refer the reader to Feferman (1991).

the statement 'All theorems of K are believable' should be accepted once we reflect upon our mathematical practice as users of K.[30] That is, once we realise that "if we believed that φ has a proof in K and we had no independent reason to disbelieve φ, then we would be ready to accept φ"; in other words, once we appreciate that proofs in K are normally good enough for us as reasons to accept their conclusions. The resulting theory $Bel(K)$, with sentences declared believable even though they are unprovable in K, permits us then to recognise our commitments as users of K.

At this point it should be emphasised that our intended interpretation of 'B' is only *partially* captured by the axioms and rules of $Bel(K)$. Apart from the formal machinery of the believability theory, the second crucial element of the picture, external to $Bel(K)$, are the prospective bridge rules which connect believability with rational acceptance. When then, or under what conditions, are we permitted to move from '$B(\varphi)$' to 'φ should be rationally accepted'? In this paper I have advocated a modest approach. I would say that given a proof of $B(\varphi)$ in $Bel(K)$, we should rationally accept φ as long as we are not aware of any derivation of $B(\neg\varphi)$ in $Bel(K)$ or in some theory $Bel(K')$, built over a well-entrenched mathematical theory K', possibly different than K. I consider it to be a workable option, one that the deflationist can employ in practice, successfully defending himself against the objections of the critics. However, there is still a room for a more ambitious approach, aiming to explain how we 'reason about reasons' even in troublesome cases.

In general, when working without the consistency axiom (A_3) (as I think we should), we are faced with the question of how the bridge rules are to be used in contexts with contradictory statements appearing within the scope of B. One of the main issues is the problem of logical explosion. Since by (A_1) the whole of first-order logic is believable, it is easy to see that with contradictory statements in the scope of B, an arbitrary statement will become believable, provably so in our theory. Let me stress that with our weak reading of 'B', this result is not troublesome in itself. Indeed, the intuition is then that treating proofs in K as reasons normally good enough to rationally accept their conclusions, gives you a reason to accept an arbitrary φ—again, a reason which is normally good enough to warrant rational acceptance. However, with both $B(\varphi)$ and $B(\neg\varphi)$ as theorems, the situation is not normal and the bridge rules should not licence the universal transition from $B(x)$ to 'x should be rationally accepted'. Indeed, various concrete cases indicate that the problem concerns not so much the believability rules and axioms, but K itself.[31]

[30] This move would not be valid with the consistency axiom (A_3) treated as a part of our characterisation of the notion of believability. The problem is that I just cannot see why 'all theorems of K are believable' should be then rationally accepted, given that our initial notion of accepting a theory is so weak. What is it that entitles us to conclude immediately that proofs in K will always function as *compelling* reasons to accept their conclusions? What is the guarantee that nothing will go wrong—that no independent reason to reject a theorem of K will ever be provided? I cannot see a good answer to this question.

[31] An example is provided by the truth theory FS, known to be ω-inconsistent (see Halbach 2011, pp. 157–158). As it happens, the proof of ω-inconsistency of FS can be reconstructed in PA; in other words, there is a formula $\varphi(x) \in L_T$ such that PA proves that: (1) $FS \vdash \exists x\varphi(x)$, (2) $\forall x\, FS \vdash \neg\varphi(x)$. Then axiom (A_1) of $Bel(FS)$ permits us to conclude that $B\big(\exists x\varphi(x)\big)$ and $\forall x\, B\big(\neg\varphi(x)\big)$, which in turn (by GEN) gives us $B\big(\forall x\neg\varphi(x)\big)$. In this way a contradiction in the scope of B is obtained, which only reflects the troublesome properties of FS itself.

Nevertheless, in the more ambitious approach afore-mentioned, one could try to give justice to the fact that we do make great efforts to tolerate contradictions in the scope of B. Naturally, when given good reasons for a pair of contradictory statements, we try to resolve the contradiction and to reassess the reasons. However, the main observation to make would be that, in the meantime, we do not treat the believability of a contradiction as an argument which, for every sentence φ of our language, demonstrates that rational acceptance of φ is invalid. As I have said, one bridge rule will permit us to conclude that φ should be rationally accepted, given that we have $B(\varphi)$ and we are not aware of any reason to accept the negation of φ. However, blocking the move to 'we should rationally accept φ' whenever both $B(\varphi)$ and $B(\neg\varphi)$ are provable, always and without exceptions, does not seem to be a good idea because in practice such an additional rule is not treated by us as universal. This intuition could lead to a more general, possibly paraconsistent analysis of both believability and rational acceptance. Indeed, I view it as a promising line of research. Nevertheless, here I prefer to remain noncommittal about the shape of such a general theory of believability and rational acceptance.

An additional avenue for further research corresponds to the limiting condition built into our Theorem 10: namely, to the reservation that the base theory K is to be formulated in a language without 'B'. Indeed, it is easy to observe that removing this reservation makes Theorem 10 false. For a trivial example, let K be PA with the additional axiom '$B(0 = 1)$'. Then clearly the standard model of arithmetic is expandable to a model of K but just as clearly $Int_{Bel(K)}$ is inconsistent. A non-trivial illustration is provided by the lottery paradox, introduced by Kyburg (1961). Consider a fair lottery with a large number of tickets exactly one of which will win. In what follows we assume that the tickets are numbered from 1 to k. It is highly probable that ticket 1 will not win, ticket 2 will not win …and so on, separately for every single ticket up to k.

The following intuitively plausible premises then lead to a contradiction:

1. It is rational to accept propositions with very high probability assigned to them.
2. If you are aware that a given proposition is inconsistent, then it is not rational to accept it.
3. If it is rational to accept φ and it is rational to accept ψ, then it is rational to accept $\varphi \wedge \psi$.

Let $R(\varphi)$ abbreviate 'it is rational to accept φ'. We obtain:

(a) R(ticket 1 will not win) and R(ticket 2 will not win) and …R(ticket k will not win), (by 1)
(b) R(ticket 1 will not win and ticket 2 will not win and …ticket k will not win), (by 3)
(c) R(ticket 1 will win or …ticket k will win), (by the assumption that some ticket will win)
(d) $R\big($(ticket 1 will not win and …ticket k will not win) and (ticket 1 will win or …ticket k will win)$\big)$ (by 3)

However, we are aware that the last sentence in the scope of R is inconsistent and in this way a contradiction with 2 is obtained.

It should be emphasised that the lottery paradox concerns rational acceptance, not believability. For believability, premise 2 simply does not hold. Believability of φ

and ψ (separately) means, in intuitive terms, that we should rationally accept both φ and ψ assuming that we are not aware of any good reason to accept $\neg\varphi$ or $\neg\psi$. We should then rationally accept $(\varphi \wedge \psi)$, unless we are aware of a good reason to accept $\neg(\varphi \wedge \psi)$, with the point being simply that a good justification of both conjuncts is normally good enough to warrant rational acceptance of a conjunction.[32] However, if we are aware that $\neg(\varphi \wedge \psi)$ is a truth of logic, then the situation is not normal and the usual bridge rule leading to rational acceptance does not apply.

Nevertheless, steps (a)-(c) could be reproduced in $Bel(K)$ assuming that the background theory K, formulated in the language with 'B', contains as its theorems (i) all sentences 'B(ticket m will not win)' for $1 \leq m \leq k$, (ii) 'B(ticket 1 will win or …ticket k will win)'. Observe that with these assumptions about K, the rule GEN is not needed at all to derive a contradiction in the scope of B, as the mere believability of propositional logic together with (A_2) is quite enough to produce this effect.

As before, if a contradiction appears in the scope of 'B', then the bridge rules had better not block the rational acceptance of *all sentences* of our language; in other words, the transition to rational acceptance should be invalidated locally, not globally. Let me also stress that the present proposal is not meant to offer a solution to the lottery paradox. Indeed, I think that the various solutions proposed in the literature can be applied in our framework on the level of the bridge rules, with the believability theory remaining neutral in this respect.[33]

In spite of these limitations, I think that the proposed epistemic strategy indeed offers an attractive way out for the disquotationalist who is worried about the truth-theoretic weakness of his axioms. It permits him to vindicate his philosophical position in a model case of arithmetical truth, where Theorem 10 applies. There are also good prospects for moving towards a conception of a more directly self-applicable notion of believability, with a base theory K possibly being formulated in the language containing 'B'.[34]

Acknowledgements The research presented in this paper was supported by the National Science Centre, Poland (NCN), grant number 2014/13/B/HS1/02892.

[32] See Ryan (1996, p. 124) for a forceful exposition of this view. In her words, 'Apart from the fact that [a conjunction principle], along with the other principles and the lottery story, generates a paradox, this appears to be an innocent epistemic principle.' Cf. also Douven (2002, p. 394), where a more general closure principle is described as one that 'we rely on in many of our everyday deliberations'.

[33] For example, if someone accepts premises 1 and 2 of the reasoning leading to the lottery paradox while claiming, with Kyburg, that rational acceptance does not agglomerate (in other words, that you cannot derive '$\varphi \wedge \psi$ should be rationally accepted' from 'φ should be rationally accepted' and 'ψ should be rationally accepted'), the proposed bridge rules would licence the transition to rational acceptance of all statements of the form 'ticket m will not win' while blocking the rational acceptance of their conjunction.

[34] Apart from introducing bridge rules for handling contradictions, another possible move would consist in iterating the application of the believability axioms and rules. For illustration, let K_0 be PA; define K_{n+1} as $Int_{Bel^{Con}(K_n)}$. We conjecture that a safety result similar to Theorem 10 can be obtained for all theories K_n in this sequence.

Appendix

In the proof of Theorem 10, the following two hierarchies will be used.

Definition 16

- $S_0 = KB \cup$ Axioms (A_1)-(A_3) of $Bel^{Con}(K)$,
- $S_{n+1} = S_n \cup \{B(\psi) : S_n \vdash \psi\} \cup \{\forall x \psi(x) : S_n \vdash \forall x\ B\psi(x)\}$,
- $S_\omega = \bigcup_{n \in N} S_n$.

We emphasise that the S-sets will be treated here as *theories* (that is, as closed under first-order consequence). Accordingly, the intended reading is that the set (say) S_1 contains everything which can be proved from S_0 with the help of additional axioms of the form '$B(\psi)$' and '$\forall x \psi(x)$', satisfying the appropriate conditions from Definition 16.

Definition 17 For a model N^* of K, we define:

- $B_0 = KB$,
- $B_{n+1} = \{\psi : \forall Z \supseteq B_n[$ if $(N^*, Z) \models (A_2) \wedge (A_3),$ then $(N^*, Z) \models \psi]\}$,
- $B_\omega = \bigcup_{n \in N} B_n$.

The proof of Theorem 10 consists in showing that: (1) $Int_{Bel^{Con}(K)} \subseteq S_\omega$; (2) $(N^*, B_\omega) \models S_\omega$. Then it follows immediately that a certain expansion of N^*, namely, (N^*, B_ω), is a model of $Int_{Bel^{Con}(K)}$.
Below we provide some details, starting with the following observation:

Observation 18 $\forall n\ B_n \subseteq B_{n+1}$.

This can be easily proved by induction. In particular, for $n = 0$, it is enough to observe that theorems of KB (that is, elements of B_0) are true in all structures (N^*, Z), independently of the choice of Z.[35]

It is also very easy to verify that:

Observation 19 $\forall n\ (N^*, B_n) \models (A_1) - (A_3)$.

The content of the next fact is that all the S-sets are contained in the corresponding sets from the B-hierarchy.

Fact 20 $\forall n\ S_n \subseteq B_{n+1}$.

Proof It is easy to observe that $S_0 \subseteq B_1$.[36] For the inductive part, assuming that $S_n \subseteq B_{n+1}$, we show that $S_{n+1} \subseteq B_{n+2}$. Fix $\psi \in S_{n+1}$ and let $(\alpha_1 \ldots \alpha_s)$ be a proof of ψ from the axioms of S_{n+1}.[37] We are going to show that $\forall k \leq s\ \alpha_k \in B_{n+2}$.

[35] We remind the reader that KB has been defined as K (a theory in the language L_K, not containing 'B'), but formulated in the language $L_{K,B}$ which is an extension of L_K with a new one place predicate 'B'. To be more precise, this means that the only axioms of KB employing the new predicate '$B(x)$' are logical axioms. In effect, an arbitrary interpretation Z of 'B' will make them true.

[36] Trivially, $KB \subseteq B_1$; it is also immediate that (A_1)–(A_3) remain true in every expansion (N^*, Z) given that $Z \supseteq B_0$ and $(N^*, Z) \models (A_2) \wedge (A_3)$. Finally, the axioms of induction for $L_{K,B}$ will remain true independently of the choice of Z.

[37] See the remark immediately below Definition 16.

Fix $k \leq s$ and assume that $\forall i < k\ \alpha_i \in B_{n+2}$. We obtain the appropriate conclusion for α_k by considering the following cases.
Case 1 $\alpha_k \in S_n$—then by the inductive assumption and Observation 18, $\alpha_k \in B_{n+2}$.
Case 2 $\alpha_k = B(\psi)$ for $\psi \in S_n$. Then by the inductive assumption $\psi \in B_{n+1}$, therefore $\forall Z \supseteq B_{n+1}(N^*, Z) \models B(\psi)$. So $B(\psi) \in B_{n+2}$.
Case 3 $\alpha_k = \forall x \psi(x)$ and $S_n \vdash \forall x\ B\psi(x)$. In this case, by the inductive assumption $\forall x\ B\psi(x) \in B_{n+1}$, therefore:

$$\forall Z \supseteq B_n[(N^*, Z) \models (A_2) \wedge (A_3) \rightarrow (N^*, Z) \models \forall x\ B\psi(x)].$$

In effect:

$$\forall Z \supseteq B_n[(N^*, Z) \models (A_2) \wedge (A_3) \rightarrow \forall x\ (\psi(x) \in Z)].$$

By Observation 19, it immediately follows that

$$\forall x\ (\psi(x) \in B_n).$$

Now we will consider two subcases:
(a) $n = 0$. Then $\forall x\ (\psi(x) \in KB)$, so $\forall Z \supseteq B_n[(N^*, Z) \models \forall x \psi(x)]$. Therefore $\forall x \psi(x) \in B_1$ (that is, to B_{n+1}, which is a subset of B_{n+2}).
(b) $n = l + 1$. Then $\forall x \forall Z \supseteq B_l[(N^*, Z) \models (A_2) \wedge (A_3) \rightarrow (N^*, Z) \models \psi(x)]$. In effect, $\forall Z \supseteq B_l[(N^*, Z) \models (A_2) \wedge (A_3) \rightarrow (N^*, Z) \models \forall x \psi(x)]$ and this means that $\forall x \psi(x) \in B_{l+1}$ (that is, it belongs already to B_n).
Case 4 α_k is obtained in the proof from α_i, α_j, with $i, j < k$ and $\alpha_j = \ulcorner \alpha_i \rightarrow \alpha_k \urcorner$. But then, by the inductive assumption, $\alpha_i, \alpha_j \in B_{n+2}$, so α_k belongs to B_{n+2} as well. □

The conclusion is that the B_n-sets provide natural models for all the S_n-s.

Corollary 21 $\forall n (N^*, B_n) \models S_n$.

Proof Fix n and $\psi \in S_n$. Since $S_n \subseteq B_{n+1}$ (Fact 20), we have: $\forall Z \supseteq B_n$ $\big($if $(N^*, Z) \models (A_2) \wedge (A_3)$, then $(N^*, Z) \models \psi\big)$. But $(N^*, B_n) \models (A_2) \wedge (A_3)$ (Observation 19); therefore $(N^*, B_n) \models \psi$. □

The next two observations can be established by induction on the length of the proof of an arbitrary ψ in $Bel^{Con}(K)$.

Fact 22 $Bel^{Con}(K) \subseteq S_\omega$.

Fact 23 $(N^*, S_\omega) \models Bel^{Con}(K)$.

As a result, we obtain the corollary stating that the interior of $Bel^{Con}(K)$ is a subset of S_ω.

Corollary 24 $Int_{Bel^{Con}(K)} \subseteq S_\omega$.

Proof Assume that $Bel^{Con}(K) \vdash B(\psi)$ (in other words, we assume that $\psi \in Int_{Bel^{Con}(K)}$). Then by Fact 23, $(N^*, S_\omega) \models B(\psi)$, therefore $\psi \in S_\omega$. □

At this stage we are able to obtain a model for all sentences which are, provably in $Bel^{Con}(K)$, in the scope of the believability predicate.

Lemma 25 $(N^*, B_\omega) \models Int_{Bel^{Con}(K)}$.

Proof Since $Int_{Bel^{Con}(K)} \subseteq S_\omega$ (Corollary 24), it is enough to show that $(N^*, B_\omega) \models S_\omega$. Fixing $\psi \in S_\omega$, we are going to prove that $(N^*, B_\omega) \models \psi$. By assumption, there is a natural number n such that $\psi \in S_n$. But $S_n \subseteq B_{n+1}$ (Fact 20), therefore $\psi \in B_{n+1}$. By definition of B_{n+1}, this means that:

$$\forall Z \supseteq B_n[\text{if } (N^*, Z) \models (A_2) \wedge (A_3), \text{ then } (N^*, Z) \models \psi].$$

Since $B_\omega \supseteq B_n$ and $(N^*, B_\omega) \models (A_2) \wedge (A_3)$,[38] we immediately conclude that $(N^*, B_\omega) \models \psi$. □

Lemma 25 permits us to obtain Theorem 10 as a direct corollary. Given a model N^* of K, we can indeed expand it to the model (N^*, B_ω), which satisfies $Int_{Bel^{Con}(K)}$.

Proceeding now to Theorem 11, we present below the proofs of two chosen cases. We are going to show that the believability of compositional axioms for negation and for the existential quantifier is derivable in $Bel(TB^-)$. Before we start, let us emphasise one technical point: if our base theory contains Peano arithmetic, we can apply GEN also in cases with more than one initial universal quantifier. For example, given '$\forall xyB(\varphi(x, y))$' (with two quantifiers) as a theorem, we are allowed to conclude '$B(\forall xy(\varphi(x, y)))$'. The reason is that in PA (and therefore also in the scope of B) we can use the pairing function freely and so we can always get rid of additional quantifiers.

Proof of Theorem 11 (chosen cases). In what follows we assume that quantifiers of the form '$\forall\psi$' and '$\exists\psi$' are restricted to arithmetical sentences (in effect, the reading is 'for every (some) arithmetical sentence ψ'). For the compositional truth axiom for negation, we claim that:

$$Bel(TB^-) \vdash B\big(\forall\psi[T(\neg\psi) \equiv \neg T(\psi)]\big).$$

The reasoning (carried out in $Bel(TB^-)$) goes as follows:

(1) $\forall\psi\ TB^- \vdash T(\neg\psi) \equiv \neg T(\psi)$ (provable in PA)
(2) $\forall\psi\ B\big(T(\neg\psi) \equiv \neg T(\psi)\big)$ (axiom (A_1))
(3) $B\big(\forall\psi[T(\neg\psi) \equiv \neg T(\psi)]\big)$. (by GEN)

Analysing the case of the existential quantifier, we are going to show that:

$$Bel(TB^-) \vdash B\big(\forall\varphi\forall a(\ulcorner\exists a\varphi\urcorner \in Sent_{L_{PA}} \rightarrow T(\exists a\varphi) \equiv \exists xT(\varphi(x)))\big).$$

We start with presenting a proof of the following fact (with 'Var' denoting the set of variables):

$$(*)\ \ Bel(TB^-) \vdash \forall\varphi(x)\forall a \in Var\, B\big(\forall a(T(\varphi(a)) \equiv \varphi(a))\big).$$

The argument for (*) (carried out in $Bel(TB^-)$) is given below. Let $E_n(x)$ be a Σ_1 arithmetical predicate with the intuitive reading 'x is a formula of syntactic complexity not larger than n'. The expression 'Tr_n' abbreviates a partial truth predicate for E_n formulas.

[38] The second conjunct follows easily from Observation 19.

(1) $\forall\varphi(x)\forall n\forall x\ TB^- \vdash \varphi \in E_n \rightarrow T(\varphi(x)) \equiv Tr_n(\varphi(x))$
(2) $\forall\varphi(x)\forall n\forall x\ B\big(\varphi \in E_n \rightarrow T(\varphi(x)) \equiv Tr_n(\varphi(x))\big)$
(3) $B\big(\forall\varphi(x)\forall n\forall x\ [\varphi \in E_n \rightarrow T(\varphi(x)) \equiv Tr_n(\varphi(x))]\big)$
(4) $\forall\varphi(x)\forall n B\big(\forall x[\varphi \in E_n \rightarrow T(\varphi(x)) \equiv Tr_n(\varphi(x))]\big)$
(5) $\forall\varphi(x)\forall n[\varphi \in E_n \rightarrow B\big(\forall x[T(\varphi(x)) \equiv Tr_n(\varphi(x))]\big)]$
(6) $\forall\varphi(x)\forall n\big(\varphi \in E_n \rightarrow TB^- \vdash \forall x[Tr_n(\varphi(x)) \equiv \varphi(x)]\big)$
(7) $\forall\varphi(x)\forall n\big(\varphi \in E_n \rightarrow B(\forall x[Tr_n(\varphi(x)) \equiv \varphi(x)])\big)$
(8) $\forall\varphi(x)\forall n\big(\varphi \in E_n \rightarrow B(\forall x[T(\varphi(x)) \equiv \varphi(x)])\big)$
(9) $\forall\varphi(x)B(\forall x[T(\varphi(x)) \equiv \varphi(x)])\big)$
(10) $\forall\varphi(x)\forall a \in Var B(\forall a[T(\varphi(a)) \equiv \varphi(a)])\big)$

(1) is provable already in PA; (2) follows from (1) by axiom (A_1); (3) is obtained by GEN. For (4), use the general statement '$\forall\alpha(x)[B(\forall x\alpha(x)) \rightarrow \forall x B(\alpha(x))]$', which is a theorem of $Bel(TB^-)$.[39] (5) uses formalised Σ_1-completeness: if $\varphi \in \Sigma_n$, then (since '$\varphi \in \Sigma_n$' is Σ_1) PA proves this fact, therefore $B(\varphi \in \Sigma_n)$ and the conclusion follows from (4). (6) is provable already in PA; (7) follows from (6) by (A_1). Step (8) is obtained from (7) and (5). (9) follows from (8) together with the information that every formula is in E_n for some n; finally, (10) is obtained from (9) together with (A_1).[40]

Given (*), we argue again in $Bel(TB^-)$ obtaining in the final move the compositional axiom for the existential quantifier.

(11) $\forall\varphi(x)\forall a \in Var\ B(\exists aT(\varphi(a)) \equiv \exists a\varphi(a)])\big)$
(12) $\forall\varphi(x)\forall a \in Var\ B(T(\exists a\varphi(a)) \equiv \exists a\varphi(a)])\big)$
(13) $\forall\varphi(x)\forall a \in Var\ B(T(\exists a\varphi(a)) \equiv \exists aT(\varphi(a))])\big)$
(14) $\forall\varphi(x)\forall a \in Var\ B(T(\exists a\varphi(a)) \equiv \exists xT(\varphi(x))])\big)$
(15) $B\big(\forall\varphi(x)\forall a \in Var\ T(\exists a\varphi(a)) \equiv \exists xT(\varphi(x))])\big)$

(11) follows from (*) by (A_1) (the believability of logic); (12) holds because TB^- is believable. (13) follows by (A_1) from (11) and (12). Step (14) involves variable renaming (a purely logical move); finally (15) is obtained by GEN. □

The proof of Theorem 15 uses the ideas which are very similar to those applied in the proof of Theorem 11.

References

Armour-Garb, B. (2010). Horwichian minimalism and the generalization problem. *Analysis*, *70*, 693–703.
Beklemishev, L. D. (2005). Reflection principles and provability algebras in formal arithmetic. *Russian Mathematical Surveys*, *60*, 197–268.
Cieśliński, C. (2011). T-equivalences for positive sentences. *Review of Symbolic Logic*, *4*(2), 319–325.
Cieśliński, C. (2015). The innocence of truth. *Dialectica*, *69*(1), 61–85.
Douven, I. (2002). A new solution to the paradoxes of rational acceptability. *The British Journal for the Philosophy of Science*, *53*(3), 391–410.
Feferman, S. (1984). Toward useful type-free theories I. *The Journal of Symbolic Logic*, *49*(01), 75–111.

[39] Assume that $B(\forall x\alpha(x))$; then fix x. We claim that $B(\alpha(x))$. By (A_1), we have: $B(\forall x\alpha(x) \rightarrow \alpha(x))$, in effect $B(\alpha(x))$ follows by (A_2).

[40] The move from (9) to (10) involves renaming the variables in the scope of B. However, this is a purely logical move and by (A_1) logic is believable.

Feferman, S. (1991). Reflecting on incompleteness. *Journal of Symbolic Logic*, *56*(1), 1–49.
Gupta, A. (1993). A critique of deflationism. *Philosophical Topics*, *21*, 57–81.
Halbach, V. (2001). Disquotational truth and analyticity. *Journal of Symbolic Logic*, *66*, 1959–1973.
Halbach, V. (2009). Reducing compositional to disquotational truth. *Review of Symbolic Logic*, *2*(4), 786–798.
Halbach, V. (2011). *Axiomatic theories of truth*. Cambridge: Cambridge University Press.
Horsten, L., & Leigh, G. (2015). Truth is simple. *Mind*.
Horwich, P. (1999). *Truth* (2nd ed.). New York: Clarendon Press.
Horwich, P. (2001). A defense of minimalism. *Synthese*, *126*, 149–165.
Horwich, P. (2010). *Truth-meaning-reality*. New York: Clarendon Press.
Ketland, J. (2005). Deflationism and the Gödel phenomena: Reply to Tennant. *Mind*, *114*(453), 75–88.
Kyburg, H. (1961). *Probability and the logic of rational belief*. Middletown: Wesleyan University Press.
McGee, V. (1992). Maximal consistent sets of instances of Tarski's schema (T). *Journal of Philosophical Logic*, *21*(3), 235–241.
Raatikainen, P. (2005). On Horwich's way out. *Analysis*, *65*, 175–177.
Reinhardt, W. N. (1986). Some remarks on extending and interpreting theories with a partial predicate for truth. *Journal of Philosophical Logic*, *15*(2), 219–251.
Ryan, S. (1996). The epistemic virtues of consistency. *Synthese*, *109*(2), 121–141.